US009116929B2

(12) **United States Patent**
Morris et al.

(10) **Patent No.: US 9,116,929 B2**
(45) **Date of Patent: Aug. 25, 2015**

(54) **WORKLOAD PRIORITY INFLUENCED DATA TEMPERATURE**

(75) Inventors: **John Mark Morris**, San Diego, CA (US); **Anita Richards**, San Juan Capistrano, CA (US); **Douglas P. Brown**, Rancho Santa Fe, CA (US)

(73) Assignee: **Teradata US, Inc.**, Dayton, OH (US)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 1519 days.

(21) Appl. No.: **11/716,880**

(22) Filed: **Mar. 12, 2007**

(65) **Prior Publication Data**

US 2008/0162417 A1 Jul. 3, 2008

Related U.S. Application Data

(60) Provisional application No. 60/877,767, filed on Dec. 29, 2006, provisional application No. 60/877,766, filed on Dec. 29, 2006, provisional application No. 60/877,768, filed on Dec. 29, 2006, provisional application No. 60/877,823, filed on Dec. 29, 2006.

(51) **Int. Cl.**
G06F 7/00 (2006.01)
G06F 17/30 (2006.01)

(52) **U.S. Cl.**
CPC ***G06F 17/30289*** (2013.01); ***G06F 17/30067*** (2013.01); ***G06F 17/30129*** (2013.01); ***G06F 17/30979*** (2013.01); *Y10S 707/99932* (2013.01)

(58) **Field of Classification Search**
None
See application file for complete search history.

(56) **References Cited**

U.S. PATENT DOCUMENTS

5,956,750	A *	9/1999	Yamamoto et al.	711/167
6,167,427	A *	12/2000	Rabinovich et al.	709/201
6,728,770	B1 *	4/2004	Bradford et al.	709/226
6,826,668	B1 *	11/2004	Hsu et al.	711/167
6,963,959	B2 *	11/2005	Hsu et al.	711/165
7,058,936	B2 *	6/2006	Chilimbi et al.	717/158
7,203,691	B2 *	4/2007	Ramesh et al.	707/101
7,359,890	B1 *	4/2008	Ku et al.	707/2
7,409,688	B2 *	8/2008	Garza et al.	718/105
7,475,108	B2 *	1/2009	Di Giulio et al.	709/203
7,523,285	B2 *	4/2009	Rider et al.	711/170
7,546,220	B1 *	6/2009	Patlashenko et al.	702/183
7,657,508	B2 *	2/2010	Morris et al.	707/999.002
7,702,676	B2 *	4/2010	Brown et al.	707/713
7,805,436	B2 *	9/2010	Richards et al.	707/720
8,082,234	B2 *	12/2011	Brown et al.	707/690
8,082,273	B2 *	12/2011	Brown et al.	707/782
8,099,411	B2 *	1/2012	Richards et al.	707/719
8,359,333	B2 *	1/2013	Brown et al.	707/792
8,392,404	B2 *	3/2013	Brown et al.	707/719
8,423,534	B2 *	4/2013	Burger et al.	707/718
8,775,413	B2 *	7/2014	Brown et al.	707/718
2003/0061352	A1 *	3/2003	Bohrer et al.	709/226
2004/0225631	A1 *	11/2004	Elnaffar et al.	707/1
2004/0243692	A1 *	12/2004	Arnold et al.	709/220
2005/0086195	A1 *	4/2005	Tan et al.	707/1
2005/0086263	A1 *	4/2005	Ngai et al.	707/104.1

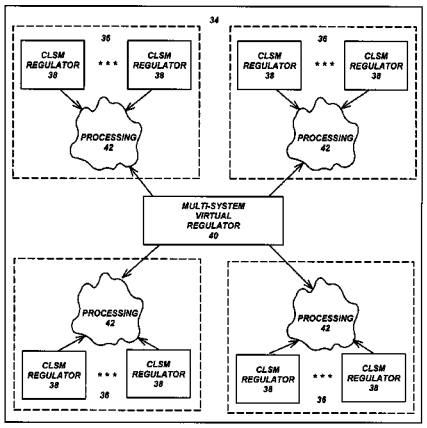
(Continued)

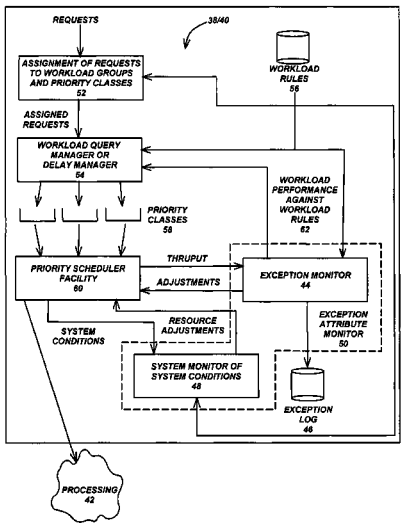
Primary Examiner — Michael Hicks
(74) *Attorney, Agent, or Firm* — Gates & Cooper LLP

(57) **ABSTRACT**

A system and method for managing one or more database systems, wherein the database systems perform database queries to retrieve data stored by the database systems. One or more regulators are used for managing the database systems, wherein the regulators monitor workload priority influenced data temperature in order to allocate resources for the systems. The data temperature is a measure of physical accesses to logical data, and the workload priority is used to further define data temperature, in order to optimize data storage placement and data access decisions.

12 Claims, 5 Drawing Sheets





(56) **References Cited**

U.S. PATENT DOCUMENTS

2005/0114862 A1* 5/2005 Bisdikian et al. 718/105
2005/0188075 A1* 8/2005 Dias et al. 709/224
2006/0218123 A1* 9/2006 Chowdhuri et al. 707/2
2007/0061375 A1* 3/2007 Brown et al. 707/200
2007/0078843 A1* 4/2007 Brown et al. 707/4
2007/0100793 A1* 5/2007 Brown et al. 707/2
2007/0174346 A1* 7/2007 Brown et al. 707/200
2008/0133456 A1* 6/2008 Richards et al. 707/2
2009/0138551 A1* 5/2009 Hubbard 709/203
2009/0327216 A1* 12/2009 Brown et al. 707/2

* cited by examiner

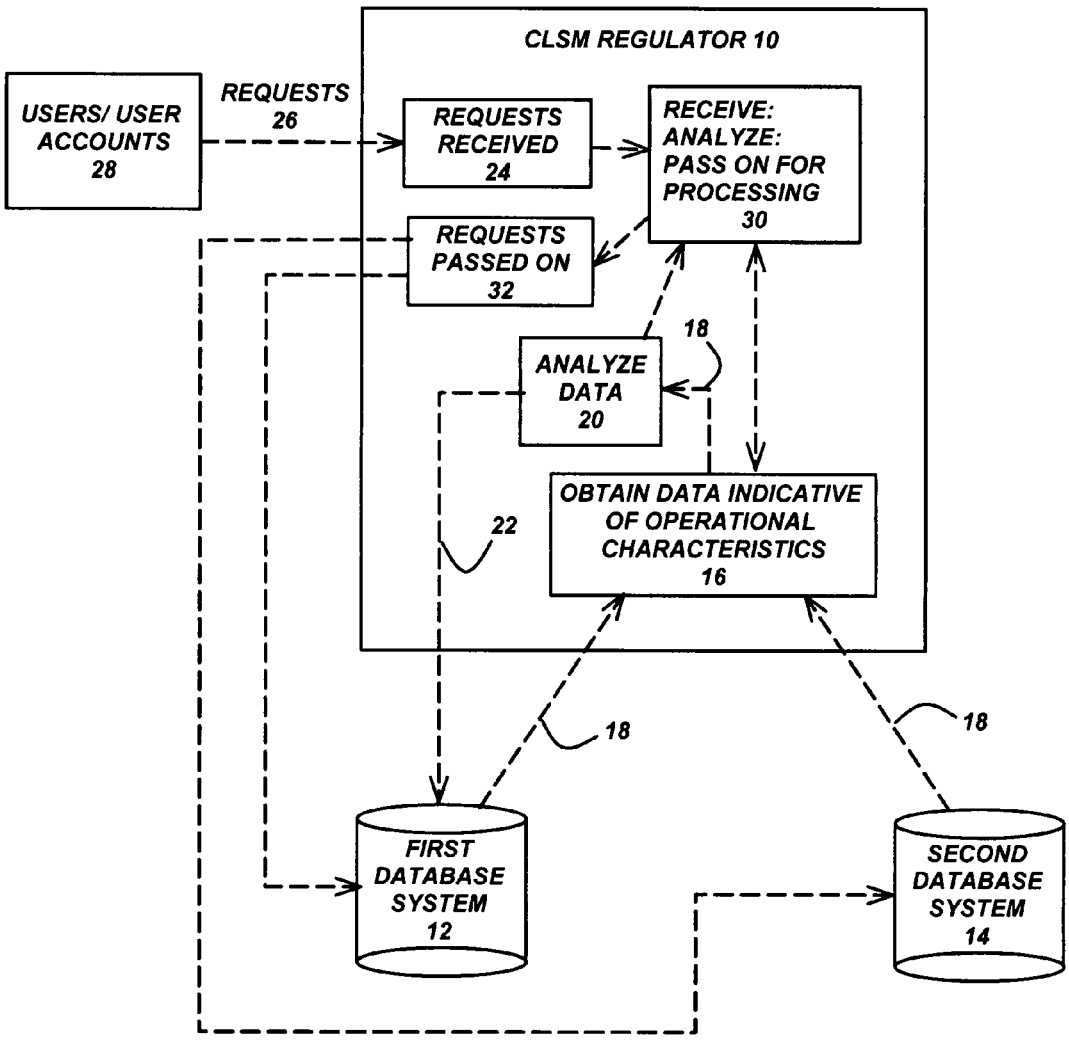

CLSM REGULATOR 10
USERS/ USER ACCOUNTS 28
REQUESTS 26
REQUESTS RECEIVED 24
RECEIVE: ANALYZE: PASS ON FOR PROCESSING 30
REQUESTS PASSED ON 32
18
ANALYZE DATA 20
OBTAIN DATA INDICATIVE OF OPERATIONAL CHARACTERISTICS 16
22
18
18
FIRST DATABASE SYSTEM 12
SECOND DATABASE SYSTEM 14

FIG. 1

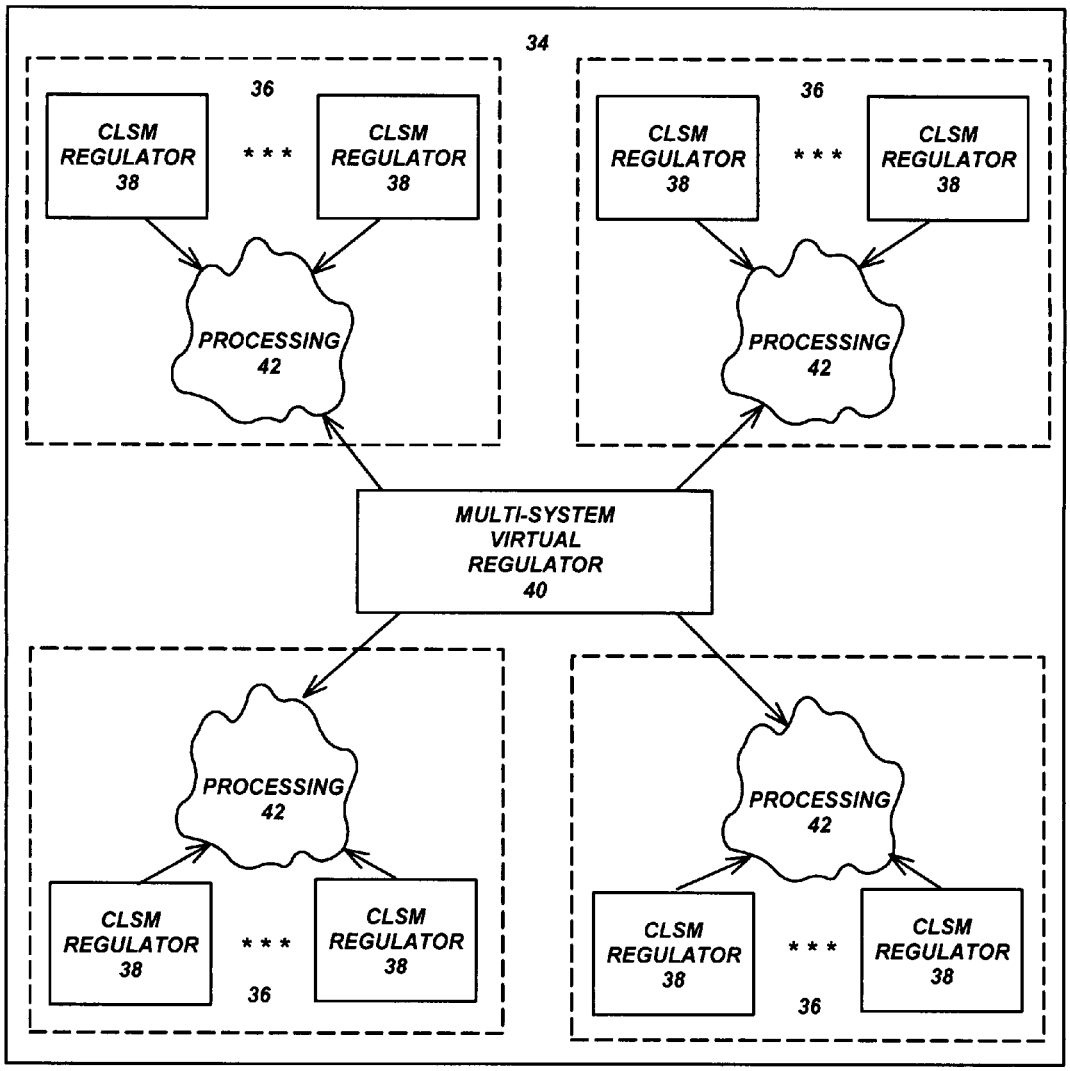

34
36
CLSM REGULATOR 38
* * *
CLSM REGULATOR 38
PROCESSING 42
36
CLSM REGULATOR 38
* * *
CLSM REGULATOR 38
PROCESSING 42
MULTI-SYSTEM VIRTUAL REGULATOR 40
PROCESSING 42
CLSM REGULATOR 38
* * *
CLSM REGULATOR 38
36
PROCESSING 42
CLSM REGULATOR 38
* * *
CLSM REGULATOR 38
36

FIG. 2

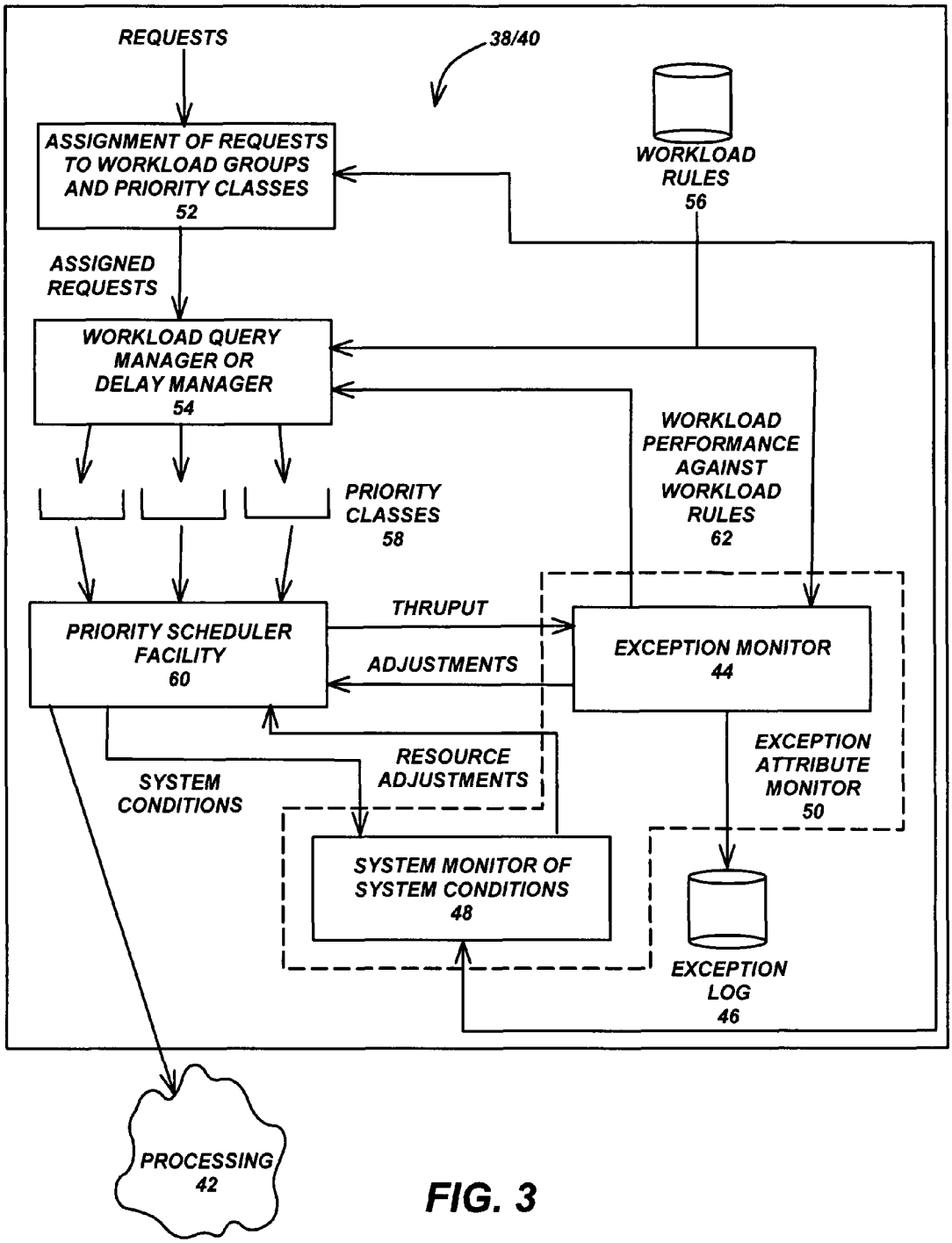

REQUESTS
38/40
ASSIGNMENT OF REQUESTS TO WORKLOAD GROUPS AND PRIORITY CLASSES 52
WORKLOAD RULES 56
ASSIGNED REQUESTS
WORKLOAD QUERY MANAGER OR DELAY MANAGER 54
WORKLOAD PERFORMANCE AGAINST WORKLOAD RULES 62
PRIORITY CLASSES 58
THRUPUT
PRIORITY SCHEDULER FACILITY 60
ADJUSTMENTS
EXCEPTION MONITOR 44
RESOURCE ADJUSTMENTS
SYSTEM CONDITIONS
EXCEPTION ATTRIBUTE MONITOR 50
SYSTEM MONITOR OF SYSTEM CONDITIONS 48
EXCEPTION LOG 46
PROCESSING 42

FIG. 3

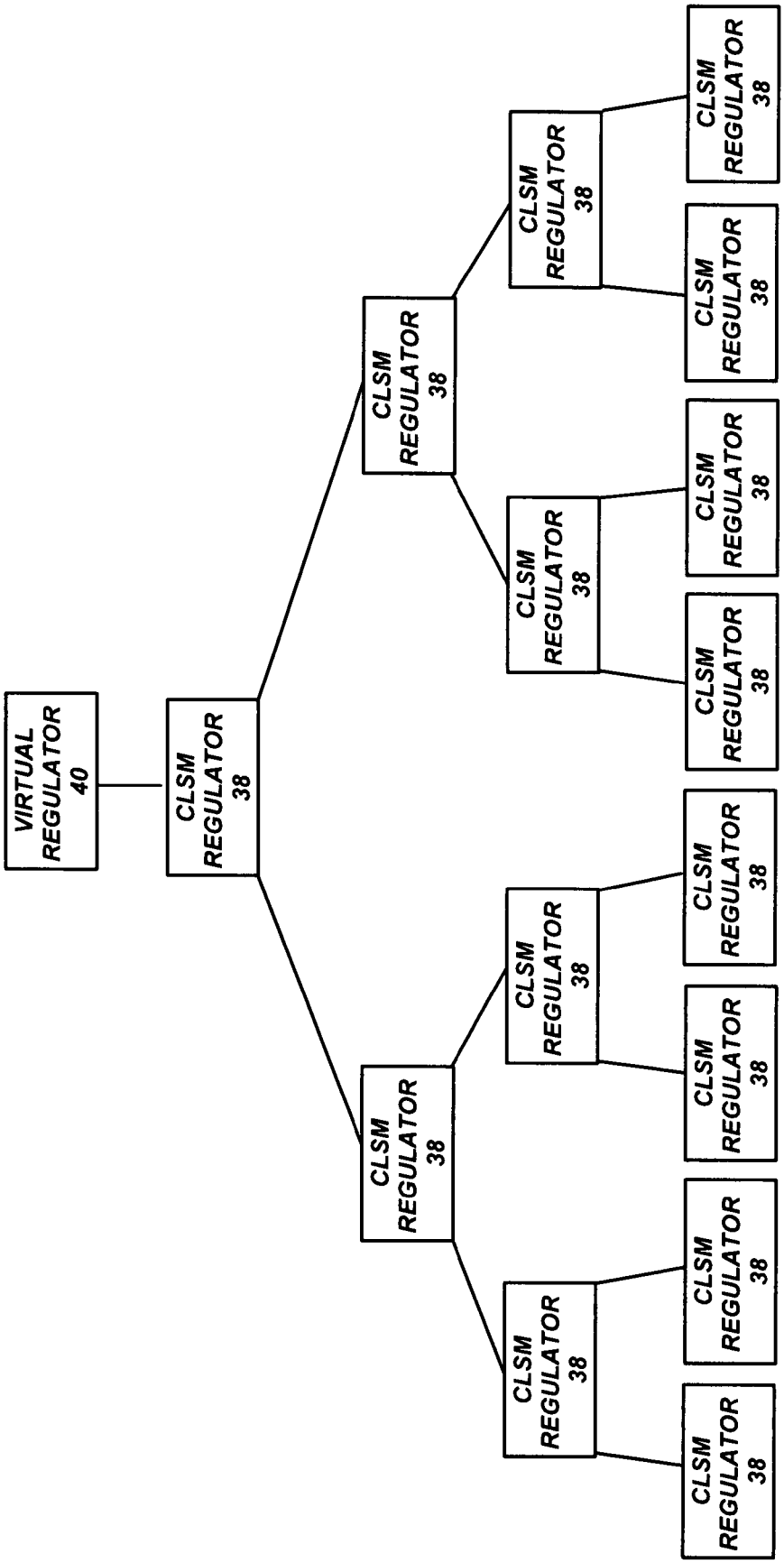

VIRTUAL REGULATOR 40
CLSM REGULATOR 38
CLSM REGULATOR 38
CLSM REGULATOR 38
CLSM REGULATOR 38
CLSM REGULATOR 38
CLSM REGULATOR 38
CLSM REGULATOR 38
CLSM REGULATOR 38
CLSM REGULATOR 38
CLSM REGULATOR 38
CLSM REGULATOR 38
CLSM REGULATOR 38
CLSM REGULATOR 38
CLSM REGULATOR 38

FIG. 4

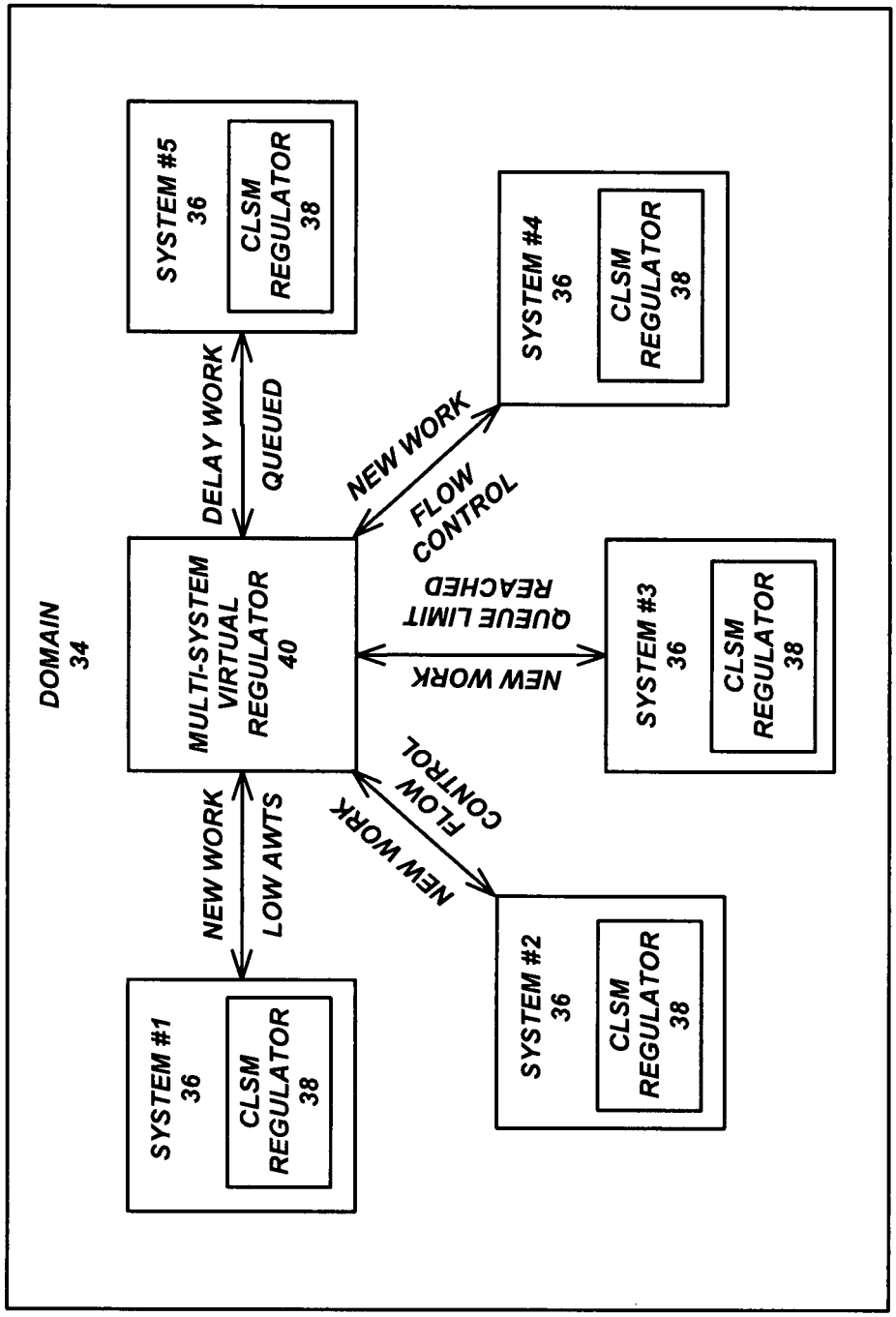

DOMAIN
34
MULTI-SYSTEM VIRTUAL REGULATOR 40
SYSTEM #1 36
CLSM REGULATOR 38
SYSTEM #2 36
CLSM REGULATOR 38
SYSTEM #3 36
CLSM REGULATOR 38
SYSTEM #4 36
CLSM REGULATOR 38
SYSTEM #5 36
CLSM REGULATOR 38
NEW WORK
LOW AWTS
NEW WORK
FLOW CONTROL
QUEUE LIMIT REACHED
NEW WORK
FLOW CONTROL
DELAY WORK
QUEUED

FIG. 5

WORKLOAD PRIORITY INFLUENCED DATA TEMPERATURE

CROSS REFERENCE TO RELATED APPLICATIONS

This application claims the benefit under 35 U.S.C. §119 (e) to the following commonly assigned applications:

U.S. Provisional Patent Application Ser. No. 60/877,767, filed on Dec. 29, 2006, by Douglas P. Brown, Anita Richards, John Mark Morris and Todd A. Walter, and entitled Virtual Regulator for Multi-Database Systems;

U.S. Provisional Patent Application Ser. No. 60/877,766, filed on Dec. 29, 2006, by Douglas P. Brown, Scott Gnau and John Mark Morris, and entitled Parallel Virtual Optimization;

U.S. Provisional Patent Application Ser. No. 60/877,768, filed on Dec. 29, 2006, by John Mark Morris, Anita Richards and Douglas P. Brown, and entitled Workload Priority Influenced Data Temperature; and

U.S. Provisional Patent Application Ser. No. 60/877,823, filed on Dec. 29, 2006, by John Mark Morris, Anita Richards and Douglas P. Brown, and entitled Automated Block Size Management for Database Objections;

all of which applications are incorporated by reference herein.

This application is related to the following co-pending and commonly assigned applications:

U.S. Utility patent application Ser. No. 10/730,348, filed Dec. 8, 2003, by Douglas P. Brown, Anita Richards, Bhashyam Ramesh, Caroline M. Ballinger and Richard D. Glick, and entitled Administering the Workload of a Database System Using Feedback;

U.S. Utility patent application Ser. No. 10/786,448, filed Feb. 25, 2004, by Douglas P. Brown, Bhashyam Ramesh and Anita Richards, and entitled Guiding the Development of Workload Group Definition Classifications;

U.S. Utility patent application Ser. No. 10/889,796, filed Jul. 13, 2004, by Douglas P. Brown, Anita Richards, and Bhashyam Ramesh, and entitled Administering Workload Groups;

U.S. Utility patent application Ser. No. 10/915,609, filed Aug. 10, 2004, by Douglas P. Brown, Anita Richards, and Bhashyam Ramesh, and entitled Regulating the Workload of a Database System;

U.S. Utility patent application Ser. No. 11/468,107, filed Aug. 29, 2006, by Douglas P. Brown and Anita Richards, and entitled A System and Method for Managing a Plurality of Database Systems, which applications claims the benefit of U.S. Provisional Patent Application Ser. No. 60/715,815, filed Sep. 9, 2005, by Douglas P. Brown and Anita Richards, and entitled A System and Method for Managing a Plurality of Database Systems;

U.S. Utility patent application Ser. No. 11/716,889, filed on Mar. 12, 2007, by Douglas P. Brown, Anita Richards, John Mark Morris and Todd A. Walter, and entitled Virtual Regulator for Multi-Database Systems;

U.S. Utility patent application Ser. No. 11/716,892, filed on Mar. 12, 2007, by Douglas P. Brown, Scott Gnau and John Mark Morris, and entitled Parallel Virtual Optimization; and

U.S. Utility patent application Ser. No. 11/716,890, filed on Mar. 12, 2007, by John Mark Morris, Anita Richards and Douglas P. Brown, and entitled Automated Block Size Management for Database Objections;

all of which applications are incorporated by reference herein.

BACKGROUND OF THE INVENTION

1. Field of the Invention

The present invention relates to a system and method for managing database systems.

2. Description of Related Art

As database management systems (DBMS) continue to increase in function and expand into new application areas, the diversity of database workloads is increasing as well. In addition to the classic relational DBMS workload consisting of short transactions running concurrently with long decision support queries, workloads comprising of an even wider range of system demands are emerging. New complex data types, such as image files, audio files, video files and other large objects, and new active data warehouse requirements, such as capacity on demand, data replication, fault-tolerance, dual active query processing, recursion, user defined types (UDFs), external UDFs, and so on, result in widely varying memory, processor, disk and network demands on database systems.

In general, a DBMS has a number of operational characteristics. These include physical statistics, such as CPU usage, query response times and performance statistics. In some DBMS, the operational characteristics include rule sets under which the database operates, relating to the likes of resource consumption and request prioritization. Varying these rule sets often has an effect on other physical characteristics, for example altering performance statistics. Ideally, a DBMS should be able to accept performance goals for a workload and dynamically adjust its own performance based on whether or not these goals are being met.

Closed-loop system management (CLSM) is a technology directed towards this ideal. Under some known CLSM-type systems, incoming queries are split into workload groups, each workload group having respective service level goals (SLGs). The DBMS is responsive to whether or not these goals are met for selectively switching between predetermined rule sets or adjusting performance controls.

It is also known to operate multi-system environments, wherein a plurality of databases, database systems, or DBMS operate in parallel. For example, DBMS that use a Massively Parallel Processing (MPP) architecture across multiple systems or a Symmetric Multiprocessing (SMP) architecture. In particular, it is known to operate a "dual-active" system wherein a plurality of databases operate in parallel and intercommunicate. It will be appreciated that managing complex workloads and performance goals performance objectives across the board in a multi-system environment is difficult.

Moreover, optimization in database system environments, especially multi-system environments, can be of critical importance. Specifically, there is a need for systems that can optimize data placement based on frequency of access and workload priority. The present invention satisfies this need.

SUMMARY OF THE INVENTION

To overcome the limitations in the prior art described above, and to overcome other limitations that will become apparent upon reading and understanding the present specification, the present invention discloses a system and method for managing one or more database systems, wherein the database systems perform database queries to retrieve data stored by the database systems. One or more regulators are used for managing the database systems, wherein the regulators monitor workload priority influenced data temperature in order to allocate resources for the systems. The data temperature is a measure of physical accesses to logical data, and the

workload priority is used to further define data temperature, in order to optimize data storage placement and data access decisions.

BRIEF DESCRIPTION OF THE DRAWINGS

Referring now to the drawings in which like reference numbers represent corresponding parts throughout:

FIG. **1** is a block diagram that illustrates the operation of a CLSM regulator according to the preferred embodiment of the present invention.

FIGS. **2**, **3**, **4** and **5** are block diagrams that illustrate the operation of a virtual regulator according to the preferred embodiment of the present invention.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENT

In the following description of the preferred embodiment, reference is made to the accompanying drawings which form a part hereof, and in which is shown by way of illustration a specific embodiment in which the invention may be practiced. It is to be understood that other embodiments may be utilized and structural changes may be made without departing from the scope of the present invention.

Overview

The present invention discloses CLSM and Virtual Regulators that use DBMS and CLSM technologies to manage workloads in a domain comprised of multiple database systems. Specifically, the Virtual Regulator has the capability to manage workloads across a domain comprised of multiple systems managed by one or more CLSM regulators in order to achieve a set of SLGs. Both the Virtual and CLSM Regulators monitor workload priority influenced data temperature in order to allocate resources for the database systems.

System Description

Referring initially to FIG. **1**, there is provided a CLSM Regulator **10** for managing a plurality of database systems **12** and **14**. The CLSM Regulator **10** includes an interface **16** for obtaining data **18** indicative of one or more operational characteristics of each of database system **12** and **14**. A monitor **20** analyzes the data **18** and provides a signal **22** indicative of an instruction to adjust one or more of the operational characteristics of a selected one of database systems **12** and **14**.

In the present disclosure, the term "database systems" is used in a general sense, and is meant to include the wider range of components used in conjunction with a database in a database system. In some embodiments, database systems **12** and **14** are simple tables of data, whereas in other embodiments they include complex relational database management systems (RDBMS). An example of such a database system **12** or **14** is the Teradata® RDBMS sold by NCR Corporation, the assignee of the present invention.

At a high level, the CLSM Regulator **10** operates as a controller including a feedback response mechanism for a domain defined by a plurality of database systems **12** and **14**. It is responsive to the performance of the domain insofar as database requests **24** are performed within predefined threshold requirements. The CLSM Regulator **10** is responsive to data **18** indicative of this performance for adjusting settings, such as resource consumption rules and query prioritization settings, in the database systems **12** and **14**. In many embodiments, this is used to better ensure that the available resources are utilized in a manner conducive to efficiently processing a variable workload.

It will be appreciated that the terms "workload class" (WC), "workload group" (WG) and "workload definition" (WD) are substantially synonymous. That is, the terms each relate to the same general identification structure used to separate requests for prioritization, processing and performance monitoring in a database system **12** or **14** including a CLSM Regulator **10**.

In the present example, the CLSM Regulator **10** analyzes the performance of each of the database systems **12** and **14** and adjusts their respective operational characteristics in response to the analysis. The analysis includes determining whether a particular class of queries are processed in accordance with one or more SLGs assigned to that class of queries.

In brief, the CLSM Regulator **10** separates incoming queries into workload groups for each database system **12** and **14** in accordance with predefined principles. Each workload group has assigned to it one or more respective SLGs. Each database system **12** and **14** maintains logs and obtains data to determine whether or not SLGs are being met for particular workload groups, and makes adjustments to operational characteristics in response. The typical objective is to adjust available settings such that the SLGs are met.

The precise nature of how workload groups are defined and settings adjusted is generally beyond the scope of the present disclosure, and various aspects are dealt with in detail in the cross-referenced applications set forth above. Other embodiments are used with database systems that use alternate architectures to analyze their performance and inherently adjust their respective operational characteristics in response to the analysis.

As noted above, in some CLSM-type systems, incoming queries are split into workload groups, wherein each workload group has its respective SLGs. The DBMS may be responsive to whether or not these goals are met for selectively switching between predetermined working values defined within a rule set or adjusting performance controls.

In this regard, the full set of workload definitions as well as filters, throttles and priority scheduler settings are considered a "rule set" for the system.

For "normal operating procedures", one set of behaviors defined for a rule set can generally manage workloads and system performance well. For example, when arrival rates per workload are in anticipated ranges, and all system components are up and functioning, a workload manager's the management of CPU, concurrency and workload exceptions can effectively manage workload performance. However, external or system-wide events can cause workload performance to stray beyond reasonable levels.

In addition, the business environment can impose restrictions on the system's performance. In the present invention, a database administrator (DBA) has the ability to specify different behaviors depending on the current situation. This is done through the definition of a two-dimensional state matrix of system conditions (SysCons) and operating environments (OpEnvs).

A sample matrix is illustrated below:

State Matrix

System	Operating Environments	
Conditions	Interactive	Batch
Green	Interactive State	Batch State
Yellow	Interactive Yellow State	Batch State

-continued

State Matrix		
System Conditions	Operating Environments: Interactive	Batch
Red	Interactive Red State	Batch State

A System Condition (SysCon) represents the "condition" or "health" of the system, e.g., degraded to the "red" system condition because a node is down.

An Operational Environment (OpEnv) represents the "kind of work" that the system is being expected to perform, e.g., within the batch operational environment, because a load job is executing.

The elements of the State Matrix are <SysCon, OpEnv> pairs. Each State Matrix element references a State. Multiple matrix elements may reference a common State. Only one State is in effect at any given time, based on the matrix element referenced by the highest severity SysCon and the highest precedence OpEnv in effect. Many <SysCon, OpEnv> pairs correlate to fewer States. Each State has a Working Value Set. A <SysCon, OpEnv> or the State can change as directed by event directives defined by the DBA.

In the present example, operational characteristics include performance statistics, rule sets (e.g., working values) under which a database system **12** or **14** is operating, physical attributes, and so on.

Some particular examples are set out below:

Memory—the amount of system and subsystem memory currently being used. It is possible that the system will include some memory that is shared among all of the subsystems.

The number of available access module processor (AMP) worker tasks (AWTs)—an AMP is a module within a database system **12** or **14** that performs a task, and an AWT is a thread or task within an AMP for performing the work assigned by a dispatcher. Each AMP has a predetermined number of AWTs in a pool available for processing. When a task is assigned to an AMP, one or more AWTs are assigned to complete the task. When the task is complete, the AWTs are released back into the pool. As an AMP is assigned tasks to perform, its available AWTs are reduced. As it completes tasks, its available AWTs are increased.

File SeGment (FSG) Cache—the amount of FSG cache that has been consumed. The FSG cache is physical memory that buffers data as it is being sent to or from the data storage facilities.

Arrival rates—the rate at which requests are arriving. Arrival rates are often broken down and used as a resource management tool on a workload basis.

Co-existence—the co-existence of multiple types of processors and or processor types.

Skew—the degree to which data (and therefore processing) is concentrated in one or more AMPs as compared to the other AMPs.

Blocking/locking—the degree to which data access are blocked or locked because other processes are accessing data.

Spool—the degree of consumption of disk space allocated to temporary storage.

Missed SLGs.

Node, cpu, memory, disk, channel, network and interconnect failures.

The CLSM Regulator **10** includes an input **24** for receiving a request **26** from a user **28**. Although user **28** is illustrated as a person, it will be appreciated that various hardware and software devices also provide requests **26**.

Request **26** is typically a database query, such as a tactical query. In the present embodiment, database systems **12** and **14** define a dual-active system where either database system **12** or **14** is capable of handling a request **26**. Despite this, it will be appreciated that one of the database systems **12** or **14** is often able to handle a request **26** more efficiently given its operational characteristics. As such, a processor **30** is responsive to interface **24** for selecting one of database systems **12** or **14**, or both database systems **12** and **14**, to process a received request **26**.

An output **32** provides the request the selected database system **12** or **14** for processing. In FIG. **1**, output **32** is shown to be providing requests to both database systems **12** and **14**. This is meant to illustrate the provision of at least two discrete requests, as well as a single request, being provided to both database systems **12** and **14**.

In the present embodiment, output **32** provides request **26** to database systems **12** and **14** in accordance with a predetermined query prioritization protocol, such as that administered by an implementation of a Priority Scheduler Facility (PSF) or a similar component. Monitor **20** adjusts this predetermined query prioritization protocol in response to data **18**. For example, in an embodiment where PSF is used, monitor **20** adjusts the PSF settings, such a class weights.

Processor **30** categorizes the request into one of a plurality of predetermined workload groups. As previously mentioned, each workload group has its respective SLGs. These SLGs relate to response times and the like, and generally comprise levels of service that are expected from database systems **12** or **14** in the processing of a request **26**. In determining which database system **12** or **14** should be selected to process a request **26**, processor **30** is responsive to operational characteristics that indicate the ability of a particular database system **12** or **14** to process a request **26** belonging to a particular workload group in accordance with the SLGs of that workload group. For example, each database system **12** and **14** is operated under one of a group of predetermined system resource consumption rules sets. Processor **30** is initiated to recognize a particular rule set as being particularly suited to handling a certain workload mix.

As a simple example, consider two generic exemplary workload groups—tactical queries and background queries. Assume that rule set A is most suitable for handling tactical queries, and rule set B is most suitable for handling background queries. For the sake of the example, interface **16** has obtained data indicative of database system **12** operating under rule set A, and database system **14** operating under rule set B. A tactical query is received by interface **24**, and recognized as a tactical query by processor **30**. Processor **30** is then responsive to interface **24** for selecting database system **12** to process that tactical query.

The above example is over simplistic to a degree. In some circumstances, interface **16** obtains other operational characteristics of database system **12** that suggest it is not meeting SLGs for tactical queries. In such a case, processor **30** selects database system **14** for the tactical query. In practical terms, tactical queries are directed to database system **14** until interface **16** obtains data to which processor **30** is responsive for altering the procedure.

Monitor **20** is responsive to processor **30** for providing a signal **22** identifying these decisions. Using the above

example, when processor **30** begins to send a stream of tactical queries to database system **14**, the workload mix of database system **14** changes. As such, rule set B is not necessarily the optimal choice; in the present example, assume that rule set C is more suitable. In such a case, monitor **20** takes the pro-active step of sending a signal **22** to database system **14** and, in response, database system **14** adapts for operation under rules set C.

Processor **30** is responsive to whether SLGs for requests **26** are being met across the domain defined collectively by database systems **12** and **14**. To this end, monitor **20** adjusts operational characteristics such as OpEnvs, SysCons, system states and/or rule sets for either or both of database systems **12** and **14**. It will be appreciated that this assists in the provision of a domain wide approach to workload administration.

It will also be appreciated that, at a high level, the CLSM Regulator **10** monitors, on a short-term basis, the execution of requests to detect a deviation from the SLGs and, where a sufficient deviation is detected, the assignment of system resources to particular workload groups across the plurality of database systems **12** and **14** are adjusted to reduce the deviation.

Referring to FIGS. **2**, **3**, **4** and **5**, embodiments will now be described with reference to a domain **34** comprised of a plurality of multiple dual-active database systems **36**, wherein each of the dual-active database systems **36** is managed by one or more CLSM Regulators **38** and the domain **34** is managed by one or more multi-system Virtual Regulators **40**.

Managing system resources on the basis of individual systems and requests does not, in general, satisfactorily manage complex workloads and SLGs across a domain **34** of database systems **36** in a multi-system environment. To automatically achieve workload goals in a multi-system environment, performance goals must first be defined (administered), then managed (regulated), and finally monitored across the entire domain (set of systems participating in an n-system environment).

CLSM Regulators **38** are used to manage workloads on an individual system **36** basis. Under the present embodiment, the Virtual Regulator **40** comprises a modified CLSM Regulator **38** implemented to enhance the CLSM architecture. That is, by extending the functionality of the CLSM Regulator **38** components, complex workloads are manageable across a domain **34**.

The function of the Virtual Regulator **40** is to control and manage existing CLSM Regulators **38** across all systems **36** in a domain **34**. The new functionality of the Virtual Regulator **38** extends the existing CLSM goal oriented workload management infrastructure, which is capable of managing various types of workloads encountered during processing **42**.

In one embodiment, the Virtual Regulator **40** includes a "thin" version of a database system **36**, where the "thin" database system **36** means a database system **36** executing in an emulation mode, such as described in U.S. Pat. Nos. 6,738,756, 7,155,428, 6,801,903 and 7,089,258. The query optimizer function of the "thin" database system **36** allows the Virtual Regulator **40** to classify received queries into "who, what, where" classification criteria, and allows the query director function of the "thin" database system **36** to perform the actual routing of the queries among multiple systems **36** in the domain **34**. In addition, the use of the "thin" database system **36** in the Virtual Regulator **40** provides a scalable architecture, open application programming interfaces (APIs), external stored procedures (XSPs), user defined functions (UDFs), message queuing, logging capabilities, rules engines, etc.

The Virtual Regulator **40** also includes a set of open APIs, known as "Traffic Cop" APIs, that provide the Virtual Regulator **40** with the ability to monitor system **36** states, to obtain system **36** status and conditions, to activate inactive systems **36**, to deactivate active systems **36**, to set workload groups, to delay queries (i.e., to control or throttle throughput), to reject queries (i.e., to filter queries), to summarize data and statistics, and to create dynamic operating rules. The Traffic Cop APIs are also made available to the CLSM Regulators **38**, thereby allowing the CLSM Regulators **38** and Virtual Regulator **40** to communicate this information between themselves in a multi-system domain **34**.

Specifically, the Virtual Regulator **40** performs the following functions:

(a) Regulate (adjust) system **36** conditions (resources, settings, PSF weights, etc.) against workload expectations (SLGs) across the domain **34**, and to direct query traffic to any of the systems **36** via a set of predefined rules.

(b) Monitor and manage system **36** conditions across the domain **34**, including adjusting or regulating response time requirements by system **36**, as well as using the Traffic Cop APIs to handle filter, throttle and/or dynamic allocation of resource weights within systems **36** and partitions so as to meet SLGs across the domain **34**.

(c) Raise an alert to a database administrator for manual handling (e.g., defer or execute query, recommendation, etc.)

(d) Cross-compare workload response time histories (via a query log) with workload SLGs across the domain **34** to determine if query gating (i.e., flow control) through altered Traffic Cop API settings presents feasible opportunities for the workload.

(e) Manage and monitor the sub-system CLSM Regulators **38** across the domain **34** using the Traffic Cop APIs, so as to avoid missing SLGs on currently executing workloads, or to allow workloads to execute the queries while missing SLGs by some predefined or proportional percentage based on shortage of resources (i.e., based on predefined rules).

(f) Route queries (traffic) to available systems **36**.

Although FIG. **2** depicts an implementation using a single Virtual Regulator **40** for the entire domain **34**, in some exemplary environments, one or more backup Virtual Regulators **40** are also provided for circumstances where the primary Virtual Regulator **40** malfunctions or is otherwise unavailable. Such backup Virtual Regulators **40** may be active or may remain dormant until needed.

Referring to FIG. **3**, both the CLSM Regulator **38** and Virtual Regulator **40** include an exception monitor **44** for detecting workload exceptions, which are recorded in a log **46**. A system condition monitor **48** is provided to detect system **36** conditions, such as node failures. These collectively define an exception attribute monitor **50**.

In practice, both the CLSM Regulator **38** and Virtual Regulator **40** receive requests, and assign these requests into their respective workload groups and priority classes at **52**. The assigned requests are then passed through a workload query manager **54**, also known as a delay manager. The workload query manager **54** is responsive to workload rules **56** and exception monitor **44** for either passing a request on or placing it in a queue until predetermined conditions are met.

If passed, the requests are split into their priority classes **58** for handling by PSF **60**. PSF **60** is responsive to the priority classes **58** for providing the requests in accordance with predefined principles for processing at **42**. These principles are updated over time in response to system monitor **48** and exception monitor **44**. PSF **60** reports observed system **36** conditions to monitor **48** and throughput information to

monitor **44**, which are responsive to such information for updating the principles under which the PSF **60** operates.

Both the CLSM Regulator **38** and Virtual Regulator **40** use a set of user-defined rules **56**, or heuristics, to guide a feedback mechanism that manages the throughput of a workload for each workload group defined in the system. In general, Virtual Regulator **40** provides a single view of managing workloads and the associated rules **56** across the domain **34**. Meanwhile, CLSM Regulators **38** continue to support workloads in a CLSM environment running on each system **36** defined in domain **34**.

The Virtual Regulator **40** manages PSF **60** settings and workload groups by controlling CLSM Regulators **38** and/or adjusting workload rules **56** in order to achieve SLGs. It also monitors operational characteristics, such as system **36** conditions, exceptions, system **36** failures, workload exceptions and the like. Further, it controls the amount of work allowed into each system **36** to meet SLGs across domain **34**.

The Virtual Regulator **40** gathers system **36** information by broadcasting to all CLSM Regulators **38** in domain **34** a request that they report their current status. This will be recognized as the functionality of interface **16** in FIG. **1**.

In some embodiments, each system **36** may have superordinate and subordinate systems **36**, and so on. An example of this is shown in FIG. **4**, which illustrates a tree structure. In such embodiments, each CLSM Regulator **38** gathers information related to its own systems **36**, as well as that of its children CLSM Regulators **38**, and reports the aggregated information to its parent CLSM Regulator **38** or the Virtual Regulator **40** at the highest level of the tree. In some cases, each CLSM Regulator **38** waits until it has received information from its children CLSM Regulators **38** before forwarding the aggregated information to its parent CLSM Regulator **38** or the Virtual Regulator **40**. In that way, the system **36** information is compiled from the bottom of the tree to the top. When the Virtual Regulator **40** compiles its information with that which is reported by all of the CLSM Regulators **38**, it will have complete information for domain **34**. The Virtual Regulator **40** analyzes the aggregated information to apply rules and make adjustments.

In the example shown in FIG. **4**, the tree is a binary tree. It will be understood that other types of trees will fall within the scope of this broad invention. Further, while the tree in FIG. **4** is symmetrical, symmetry is not a limitation.

In another example system, each CLSM Regulator **38** communicates its system **36** information directly to the Virtual Regulator **40**. The Virtual Regulator **40** compiles the information, adds domain **34** or additional system **36** level information, to the extent there is any, and makes its adjustments based on the resulting set of information.

Each CLSM Regulator **38** monitors and controls, in a closed loop fashion, workload group performance information for a single system **36** or dual-active system **36**. For example, this may require performance information received from a dispatcher processor, wherein the performance information is compared to SLGs **62**. In the example of throughput information, the level of desired throughput defined in SLGs **62** is compared to the actual level of throughput occurring for a particular workload. The Virtual Regulator **40** then adjusts resource allocation weights to better meet the workload rules.

Referring to FIG. **5**, the Virtual Regulator **40** receives information concerning the states, events and conditions of the systems **36** from the CLSM Regulators **38**, and compares these states, events and conditions to the SLGs **62**. In response, the Virtual Regulator **40** adjusts the operational characteristics of the various systems **36** through a set of "Traffic Cop" Open APIs to better address the states, events and conditions of the systems **36** throughout the domain **34**.

To manage workloads among dynamic domain **34** wide situations, the Virtual Regulator **40** classifies the various states, events and conditions into at least three general detection categories and specifies what automated actions should occur in response thereto, as set forth below:

1. Detection category: system, device or application state.
   a. This category detects the state of a system **36**, device or application (e.g., down, recovered, degraded, etc.), as communicated through a monitored message queue.
2. Detection category: system, device or application event.
   a. One event detected is an operating window time boundary.
   b. Another event detected is when an application, device or system is started or ended (as communicated through a monitored message queue). For example, in order for a request to the database system **36** to enter, initiate or continue under a given workload group, it can be additionally pre-qualified to satisfy one or more event-based "when" conditions.

   An example of the types of events supported includes "daily load against table X is about to start." This event triggers a phased set of actions: (a) begin acquisition phase of multi-load to table X, (b) promote the priority of all queries that involve table X, (c) at the same time, restrict the ability for new queries involving table X from starting until after the data load is completed (do this through delay, scheduling or disallowing the query upon request), (d) after y minutes or upon completion of the acquisition phase (which ever comes later), previously promoted queries that are still running are aborted ("times up!"), (e) begin the apply phase of the data load, and (f) upon completion of data load, raise restrictions on queries involving table X, and allow scheduled and delayed queries to resume.

   Another example is to allow the user to define and automate workload group changes based on an event (rather than resource or time changes per the PSF **60**). For example, customers may like to have workload groups change when the daily load application is submitted to the system **36**, or based on a business calendar that treats weekends and holidays differently from weekdays, or that treats normal processing differently from quarterly or month-end processing.

   What is different about this type of initiation is that it may mean one query or event can have an impact on other queries already in execution or that will soon be requested.
   c. Another event detected is sustained CPU and I/O for some qualifying time, e.g., either high or low. For example, the category may initiate background tasks when system **36** utilization is low, and eliminate them when system **36** utilization is high.
   d. User-definable event.
   e. Replication service.
   f. Other events.
3. Detection category: system, device or application condition. These conditions are detected by an individual request or workload group being impacted (e.g., average response time greater than the SLG for some qualifying interval) by a condition and can be implemented through existing exception monitoring rather than more difficult domain **34** wide detection. Preferably, there are "symptom" and "cause" conditions:
   a. Symptom: response time greater than x.
   b. Symptom: block time greater than x.

c. Cause: low or no AWT availability for some time period.

d. Cause: arrival rate greater than expected (e.g., a surge).

e. Cause: response time for all workload groups exceeded.

f. Cause: response time for one workload group exceeded.

4. Automated actions. Upon detection of any of the above states, events, or conditions, one or more automated actions can be triggered by the Virtual Regulator **40** or CLSM Regulator **38**. Automated actions may include (but are not limited to) the following:

a. Alerting an operator.

b. Notifying one or more systems **36**.

c. Logging the states, events and conditions.

d. Changing the rules for one or more workload groups.

e. Re-routing one or more workload groups to other systems.

f. Aborting one or more workload groups.

Alternatively, there may be some other action taken in order to automatically resolve the detected category. Moreover, many other categories of detections and automated actions can be implemented.

Generally speaking, CLSM Regulators **38** provide real-time closed-loop system management over resources within the systems **36**, with the loop having a fairly narrow bandwidth, typically on the order of milliseconds, seconds, or minutes. The Virtual Regulator **40** provides real-time closed-loop system management over resources within the domain **34**, with the loop having a much larger bandwidth, typically on the order of minutes, hours, or days.

Further, while CLSM Regulators **38** controls resources within the systems **36**, and the Virtual Regulator **40** controls resources across the domain **34**, in many cases, system **36** resources and domain **34** resources are the same. The Virtual Regulator **40** has a higher level view of resources within the domain **34**, because it is aware of the state of resources of all systems **36**, while each CLSM Regulator **38** is generally only aware of the state of resources within its own systems **36**.

There are a number of techniques by which Virtual Regulator **40** implements its adjustments to the allocation of system **36** resources. For example, and as illustrated in FIG. **2**, the Virtual Regulator **40** communicates adjustments directly to CLSM Regulators **38**, and the CLSM Regulators **38** then apply the relevant rule adjustments. Alternatively, the Virtual Regulator **40** communicates adjustments to the CLSM Regulators **38** by passing them down a tree, such as that in FIG. **4**. In either case, the CLSM Regulators **38** incorporate adjustments ordered by the Virtual Regulator **40** in the various systems **36**.

Given that the Virtual Regulator **40** has access to the state, event and condition information from all CLSM Regulators **38**, it can make adjustments that are mindful of meeting SLGs for various workload groups. It is capable of, for example, adjusting the resources allocated to a particular workload group on a domain **34** basis, to make sure that the SLGs for that workload group are met. It is further able to identify bottlenecks in performance and allocate resources to alleviate the bottleneck. Also, it selectively deprives resources from a workload group that is idling resources. In general, the Virtual Regulator **40** provides a domain **34** view of workload administration, while the CLSM Regulators **38** provide a system **36** view of workload administration.

Thus, it will be appreciated that the illustrated Virtual Regulator **40** is capable of monitoring the performance and operational characteristics of a plurality of systems **36** across a domain **34**. From this, it provides a domain **34** based approach to resource and performance management.

Workload Priority Influenced Data Temperature

Frequency of access is critically important to optimization of data storage placement for system **36** efficiency and performance. Looking beyond system **36** efficiency, there is an opportunity to influence data storage placement and data access according to the dynamic workload group priority.

Data temperature is commonly understood to represent frequency of access to data as a measure of the interaction, level of interest and importance of that data. Specifically, the data temperature is a measure of physical accesses to logical data, wherein "hot" data is frequently accessed, "warm" data is moderately accessed, and "cold" data is infrequently accessed.

Since the level of activity is the key determinant of data temperature, there is a need for dynamic monitoring of data temperature conditions and dynamic allocation of resources in response to the data temperature conditions. In one embodiment of the present invention, workload group priority influenced data temperature is one of the conditions monitored by the systems **36**, CLSM Regulators **38** and Virtual Regulator **40**. Moreover, this condition then affects the allocation of resources by the systems **36**, CLSM Regulators **38** and Virtual Regulator **40**.

As noted above in FIG. **3**, both the Virtual Regulator **40** and the CLSM Regulators **38** assign received requests into their respective workload groups and priority classes at **52**. The assigned requests are then passed through a workload query manager **54**, which is responsive to workload rules **56** and exception monitor **44** for either passing a request on or placing it in a queue until predetermined conditions are met.

If passed, the requests are split into their priority classes **58** for handling by PSF **60**, which is responsive to the priority classes **58** for providing the requests in accordance with predefined principles for processing at **42**. PSF **60** reports observed system **36** conditions to monitor **48** and throughput information to monitor **44**, which are responsive to such information for updating the principles under which the PSF **60** operates.

Also as noted above, both the Virtual Regulator **40** and the CLSM Regulators **38** use the set of user-defined rules **56** to guide the feedback mechanism that manages the throughput for each workload group. Moreover, the Virtual Regulator **40** manages PSF **60** settings and workload groups by controlling CLSM Regulators **38** and/or adjusting workload rules **56** in order to achieve SLGs. The Virtual Regulator **40** also monitors operational characteristics, such as system **36** conditions, exceptions, system **36** failures, workload exceptions and the like. Further, it controls the amount of work allowed into each system **36** to meet SLGs across domain **34**.

In the present invention, the priority of the workload groups are passed down to the systems **36** by the PSF **60**, so that the systems **36** can use the priority, in conjunction with data temperature, for optimization of data storage placement and/or data access. Previously, the system **36** had no way of knowing the importance of any given I/O or associated CPU. While the system **36** aims for I/O and CPU efficiency that benefits overall performance, the present invention provides an opportunity to use workload group priority to influence data temperature, and thereby data storage placement and data access, such that all I/Os are not created equal.

By tying in workload group priority to data temperature, e.g., by using workload group priority to further define data temperature, the system **36**, CLSM Regulator **38** and Virtual Regulator **40** can further optimize data storage placement and data access decisions. One example would be to rank a high priority workload group accessing an 82° cylinder (e.g., a longer seek and thus representative of "cooler" data) for a

particular disk drive over a lesser priority workload group accessing a 91° cylinder (e.g., a shorter seek and thus representative of "warmer" data) for the same disk drive.

Note that the workload group priority may be logged in order to be integrated with data temperature information. By integrating the data temperature information with workload group priority logging, it is possible to discover additional insight that can be fed back into the data storage placement.

CONCLUSION

This concludes the description of the preferred embodiment of the invention. The following describe some alternative embodiments for accomplishing the same invention.

The invention has been primarily developed for monitoring and adjusting the operational characteristics of a plurality of systems within a domain. However, it will be appreciated that the invention is in no sense limited to that application. For example, the invention is generally applicable to a wide variety of environments where such functionality has value.

The foregoing description of the preferred embodiment of the invention has been presented for the purposes of illustration and description. It is not intended to be exhaustive or to limit the invention to the precise form disclosed. Many modifications and variations are possible in light of the above teaching. It is intended that the scope of the invention be limited not by this detailed description, but rather by the claims appended hereto.

What is claimed is:

1. A system for managing a plurality of database systems, comprising:

(a) one or more computers;

(b) a domain comprised of a plurality of database systems executed by the one or more computers, wherein each of the plurality of database systems performs database queries to retrieve data stored by the plurality of database systems; and

(c) a plurality of closed-loop system management (CLSM) regulators for managing the database systems wherein:

(i) the plurality of CLSM regulators are organized in a tree structure, such that each CLSM regulator aggregates information related to its own database system as well as that of its children CLSM regulators, and reports the aggregated information to its parent CLSM regulator in the tree structure;

(ii) at least one of the plurality of CLSM regulators comprises a virtual regulator at a highest level of the tree structure for managing the domain by controlling the CLSM regulators managing the database systems within the domain, and communicating with the CLSM regulators managing the database systems in the domain to compile the aggregated information reported by the CLSM regulators;

(iii) each of the plurality of CLSM regulators dynamically monitor a data temperature comprising a frequency of access to the data stored by a specific database system;

(iv) each of the plurality of CLSM regulators separate incoming queries into one or more workload groups;

(v) each of the one or more workload groups is assigned to one or more respective service level goals (SLGs) that are each comprised of one or more levels of service expected from the plurality of database systems in processing the incoming queries assigned to the respective workload group;

(vi) each of the plurality of CLSM regulators dynamically determine a priority class for each workload group based on the respective service level goals;

(vii) a workload query manager within each of the plurality CLSM regulators is responsive to workload rules and an exception monitor for either passing on the incoming query or placing the incoming query into a queue until predetermined conditions are met, wherein once passed, the incoming query is placed into its determined priority class;

(viii) a priority scheduler facility (PSF) within each of the plurality of the CLSM regulators is responsive to the priority classes, and reports observed system conditions and throughput information to one or more monitors;

(ix) the one or more monitors update principles under which the PSF operates based on the observed system conditions and the throughput information, received from the PSF, such that the priority class influences and is used to define the data temperature; and

(x) each of the plurality of the CLSM regulators dynamically utilize the priority in conjunction with the data temperature to dynamically allocate resources for the database systems for processing the one or more workload groups.

2. The system of claim **1**, wherein the data temperature is a measure of physical accesses to logical data.

3. The system of claim **1**, wherein the workload priority is used for data storage placement optimization.

4. The system of claim **1**, wherein data storage placement in the systems is influenced according to a priority of a workload group's accesses to the data.

5. The system of claim **1**, wherein the workload priority is used for data access optimization.

6. The system of claim **1**, wherein data access in the systems is influenced according to a priority of a workload group's accesses to the data.

7. A method for managing database systems, comprising:

(a) executing a plurality of database systems, wherein each of the plurality of database systems performs database queries to retrieve data stored by the plurality of database systems; and

(b) managing the database systems using a plurality of closed-loop system management (CLSM) regulators wherein:

(i) the plurality of CLSM regulators are organized in a tree structure, such that each CLSM regulator aggregates information related to its own database system as well as that of its children CLSM regulators, and reports the aggregated information to its parent CLSM regulator in the tree structure;

(ii) at least one of the plurality of CLSM regulators comprises a virtual regulator at a highest level of the tree structure for managing the domain by controlling the CLSM regulators managing the database systems within the domain, and communicating with the CLSM regulators managing the database systems in the domain to compile the aggregated information reported by the CLSM regulators;

(iii) each of the plurality of CLSM regulators dynamically monitor a data temperature comprising a frequency of access to the data stored by a specific database system;

(iv) each of the plurality of CLSM regulators separate incoming queries into one or more workload groups;

(v) each of the one or more workload groups is assigned to one or more respective service level goals (SLGs)

that are each comprised of one or more levels of service expected from the plurality of database systems in processing the incoming queries assigned to the respective workload group;

(vi) each of the plurality of CLSM regulators dynamically determine a priority class for each workload group based on the respective service level goals;

(vii) a workload query manager within each of the plurality CLSM regulators is responsive to workload rules and an exception monitor for either passing on the incoming query or placing the incoming query into a queue until predetermined conditions are met, wherein once passed, the incoming query is placed into its determined priority class;

(viii) a priority scheduler facility (PSF) within each of the plurality of the CLSM regulators is responsive to the priority classes, and reports observed system conditions and throughput information to one or more monitors;

(ix) the one or more monitors update principles under which the PSF operates based on the observed system conditions and the throughput information, received from the PSF, such that the priority class influences and is used to define the data temperature; and

(x) each of the plurality of the CLSM regulators dynamically utilize the priority in conjunction with the data temperature to dynamically allocate resources for the database systems for processing the one or more workload groups.

8. The method of claim **7**, wherein the data temperature is a measure of physical accesses to logical data.

9. The method of claim **7**, wherein the workload priority is used for data storage placement optimization.

10. The method of claim **7**, wherein data storage placement in the systems is influenced according to a priority of a workload group's accesses to the data.

11. The method of claim **7**, wherein the workload priority is used for data access optimization.

12. The method of claim **7**, wherein data access in the systems is influenced according to a priority of a workload group's accesses to the data.

* * * * *